Patent Office
Research and Development
Reports . . . . . No. 8

RECENT ADVANCES IN PATENT OFFICE SEARCHING:

STEROID COMPOUNDS
AND
ILAS

FOR OFFICIAL DISTRIBUTION

STEROID COMPOUNDS

Prepared by

Julius Frome
H. R. Koller
Jacob Leibowitz
H. Pfeffer

Patent Research Specialists
Staff Members

ILAS

Prepared by

Don D. Andrews, *Director*

Office of Research and Development

U. S. Patent Office

Robert C. Watson
Commissioner of Patents

Sinclair Weeks
Secretary of Commerce

RECENT ADVANCES IN PATENT OFFICE SEARCHING

STEROID COMPOUNDS

Preface

This is the first part of a set of two reports embodying the presentation made at the Western Reserve University Symposium for Systems on Information Retrieval held in Cleveland, Ohio, on April 15-17, 1957.

INTRODUCTION

The U. S. Patent Office and the National Bureau of Standards are engaged in a joint research program to develop and apply automatic techniques of literature searching toward the solution of the Patent Office search problems.

In the course of examination of an application for patent the Patent Examiner conducts what is known as a prior art search. The purpose of this search is to find the most pertinent antecedent art to which the application at hand relates on the basis of which the Examiner is enabled to make a determination of patentability.

SEARCH PROBLEMS

There are several characteristics of a prior art patent search which are of significance in consideration of the Patent Office search problems.

First, since the search is made from the point of view of patentability, the test for the degree of pertinence of the subject matter of search to the subject matter of the application for patent is governed by the established criteria for patentability. Hence, the search is not only for disclosures of identical concepts but also for concepts which are analogous and similar according to these criteria. To find such similiarities, the search is made, in effect, on the basis of a class or genus which includes both the specified subject matter of the application and all related subject matter. This is called generic searching.

Secondly, since each application is directed, *prima facie*, to novel and inventive subject matter, the points of view of search are both extremely variable and not readily susceptible to precognition. There is a need, therefore, for great versatility in search terminology and the ability to search according to such terminology. Provision must also be made for the ability to conduct the search through the vast subject matter field of science and technology encompassed by the Patent Office.

Thirdly, with respect to any search, there is ordinarily no foreknowledge as to the existence or absence of the desired information. The existence of the information is evidenced by finding it, while failure to find it leads to a presumption that it is absent, which presumption will govern the Examiner's action on the application. However, to give validity to such a presumption it is necessary to remove another possible cause for failure to find the information, namely an ineffective search mechanism.

Because automatic data processing techniques appear to offer the best means of arriving at an effective solution to this problem, the Patent Office, through its Office of Research and Development, is undertaking the mechanization of the Patent Office search. With respect to the chemical arts, a routine has been worked out for performing comprehensive chemical searches, which includes methods for searching with respect to chemical products, products of natural origin, processes, functions and various chemical interrelationships and correlations. This routine is now being tested on SEAC at the National Bureau of Standards.

CHEMICAL COMPOUND SEARCH PROBLEM

Our description here is confined to the subject matter of chemical compounds per se, illustrating several different mechanization techniques.

With chemical compounds, as with other subject matter, the search is generic for two major reasons. First, since the claims (which define the alleged invention) of the patent application may specify a whole class of compounds, any member of which if already known would bar the allowance of the class, the search must be for any and all members of the genus defined by the claims. Secondly, if the specific compound or class of compounds claimed are new, the Examiner must still investigate analogous compounds to determine whether the differences between the old and the new are sufficient to warrant the grant of a patent. Since

compounds are analogous to each other on the basis of common characteristics, the search would be for the genus of compounds having these characteristics.

To illustrate, in the skeletal structural diagrams of the compounds illustrated in figure 1—

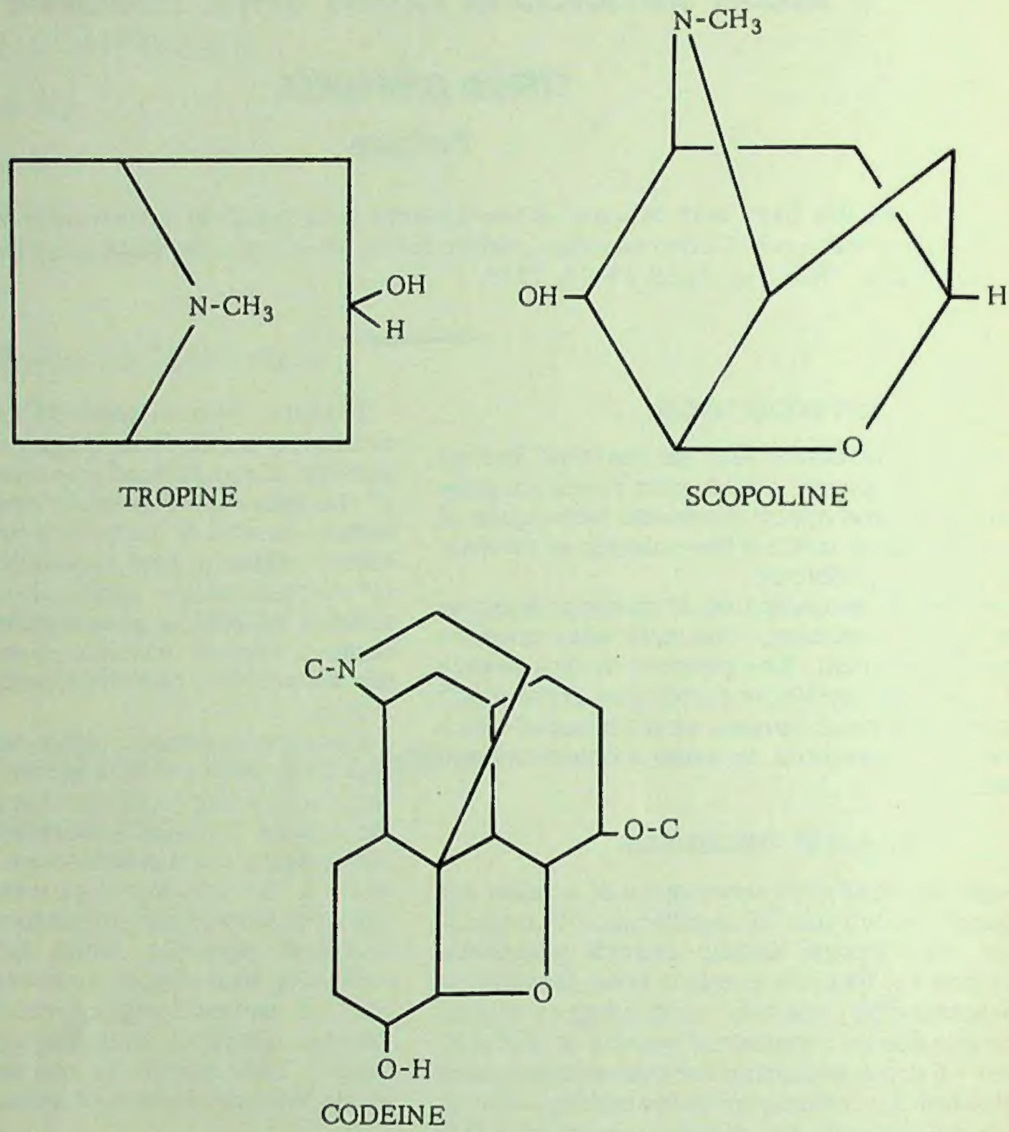


Figure 1.

tropine, scopoline and codeine are members of the class of compounds containing a six membered N ring. Note that the 6 membered N ring fragment is present in these three structures despite the distortion in the portrayal. From the point of view of 5 membered N ring compounds, tropine and scopoline are members of the same class with codeine excluded. On the other hand, from the point of view of 5 membered O ring compounds, scopoline and codeine are of the same class, with tropine excluded. Thus, the compounds are collected and separated according to the particular class basis under consideration. A search based on these class terms is expected to collect and separate according to the requirements posed by the terminology of the search request.

These compounds have been described in terms of configurations of elements contained within the more comprehensive complete structure. These substructures or "fragments" are, in effect, generic descriptors for the compounds. It is evident that the number of terms that can be derived from all possible substructure combinations of elements within the average chemical compound is fantastically high—yet, any one of these terms constitutes a potential generic search question. It is manifestly impossible, by conventional means, to establish such a list of genera and classify all compounds

according to all their possible descriptors for retrieval. Hence in conventional classification, a judicious choice is made, on a practical basis, of a number of descriptors and each compound is ordinarily classified according to one of the terms of this pre-selected list, according to rules of priority among the terms.

FIRST TOPOLOGICAL SYSTEM

The first system to be described possesses the desired capability of making a search based on any possible chemical structural configuration and permitting retrieval of any and all compounds containing the requested configuration, without the use of any pre-designated lists or descriptions. This has been termed the topological method.

The theory behind the topological method is as follows. It is known that if a chemist is presented with the structural formula diagrams of the compounds of figure 1 and is asked, "Are these hydroxy (OH) compounds?", "Are they 5 membered O ring compounds?", "Are they nitro (NO_2) compounds?", he can almost immediately answer, "yes" or "no," as the case may be. Thus, from the formula, all possible structural subgroups are potentially available as generic descriptors and each compound is considered only in terms of the descriptors involved in the current search question.

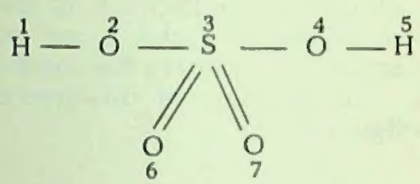
The topological search technique describes the structural formula of each disclosed compound to the computer (SEAC) in such a way that the computer can analyze both the disclosed compound and the question structure or substructure and arrive at the same determination made by the chemist in his visual inspection.

The subject matter of the first topological search is the steroid compounds of which cortisone, digitalis and the sex hormones are familiar examples.

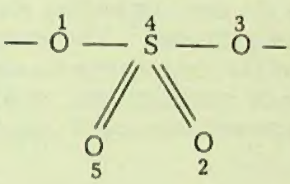
Coding

The coding rules are quite simple and can be done by non-chemically trained clerical personnel from the structural formula. To illustrate, consider the compound, sulfuric acid, in figure 2.

SULFURIC ACID



SULFATE RADICAL



CODE FOR SULFURIC ACID

No	Hex	IV	III	II	I	Element
1	0 1				0 2	0 1
2	0 2			0 3	0 1	0 8
3	0 3	0 7	0 6	0 4	0 2	1 6
4	0 4			0 5	0 3	0 8
5	0 5				0 4	0 1
6	0 6				0 3	0 8
7	0 7				0 3	0 8

CODE FOR SULFATE RADICAL

No	Hex	IV	III	II	I	Element
1	0 1				0 4	0 8
2	0 2				0 4	0 8
3	0 3				0 4	0 8
4	0 4	0 5	0 2	0 3	0 1	1 6
5	0 5				0 4	0 8

Figure 2.

1. First, each atom in the structural formula is assigned a number for identification, called a sequence number. The numbers are sequential but their assignment to the atoms is entirely random and arbitrary.

2. The compound is then coded on a coding sheet, as illustrated below the formula. The two left hand columns are for sequence numbers and their hexadecimal equivalents. The right hand column is for identification in code (atomic number) of the ele-

ments. Fields I, II, III and IV are for connectivity relationships of the elements. The coding is done for each element, row by row, starting with element 1. Thus, the first row states that element O1 is hydrogen (which has code O1) and that it is connected to an element of sequence number O2. The next row states that element O2 is oxygen (which has code O8) and that it is connected to an element of sequence number O1, and an element of sequence number O3. The coding is thus continued, each element in the compound and its connectivity to other elements being defined. From this code, the compound and any segment of the compound can be reconstructed. If the elements had been numbered differently, the same reconstruction would be obtained due to the equivalence of the connectivity pattern.

When the code is completed and stored on magnetic tape in the memory of the computer it is available for search. A searcher wishing to investigate the file as to the presence of the entire molecule or some "fragment" of it, as the sulfate radical, would code the question in a similar manner, making his own sequence number assignment. It will be noted that the elements of the sulfate radical in figure 2 do not have the same number assignment as in the corresponding elements in the sulfuric acid configuration. The computer, however, by means of its programmed instructions will perform a series of comparisons, making an atom-to-atom match between the coded question and each structure in the file to find out whether or not the question configuration is within the disclosure structure. If such a corresponding structure is found, the patent number is printed and the search proceeds to the next structure.

Results

The first test of this method on a file of 250 complex steroid compounds yielded encouraging results. Questions could be asked in terms of any desired chemical substructural configuration, without any foreknowledge of the complete structure or how it was coded. It appears that this method solves the problem of generic searching from the point of view of any genus that is capable of structural representation. Furthermore, the connectivity definition permits the finding of synonymous structures unhampered by distortion in their visual portrayals as seen in figure 1. An important aspect of the system, also, is the simplicity and uniformity of its rules.

SECOND TOPOLOGICAL SYSTEM

Following this initial test, a second experiment was launched to deal with certain problems. One problem of major significance is that concerned with the so-called artificial genus or "Markush" formula customary in patent disclosures and search questions. An example of this type of formula is shown in figure 3.

STEROID NUCLEUS

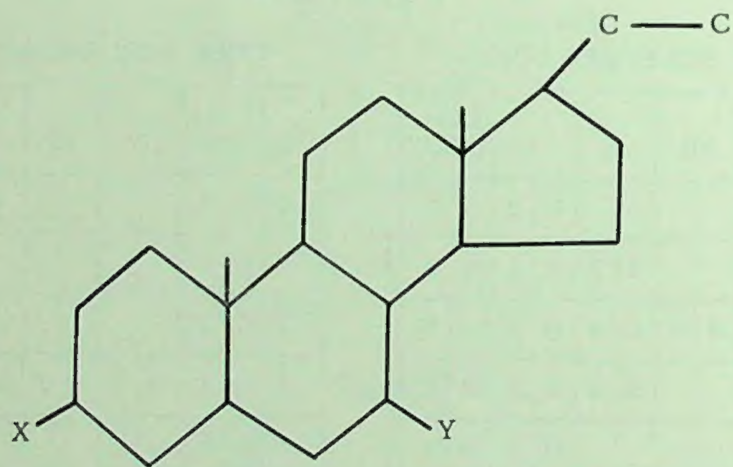


Figure 3.

X being defined as a member of the group consisting of OH, Cl, Br and Y as an O-acyl of type 1, 7, or 9 (1, 7, and 9 being further defined). It will be easily seen that, given a few such variables, literally thousands of compounds can be encompassed within a single formula representation, each of them constituting a valid disclosure of the compound concept. This type of formula can be presented in a search question as well as in the disclosure with no necessary conformity in scope between the question and the disclosure answer. It would be a matter of great difficulty to encode the vast number of theoretically possible combinations as well as to search through the resulting expanded file.

In order to concentrate on the Markush problem and to find ways and means for increasing searching speed, a limitation was imposed on the uni-

versality of the system, i.e., its ability to search regardless of the type of compounds involved. An experiment was therefore undertaken in which the search was confined to steroids, i.e., all searches required that there be at least a steroid nucleus in the requested structural group.

Coding

According to the method of the second topological experiment, the compound of figure 3 is prepared for coding as follows.

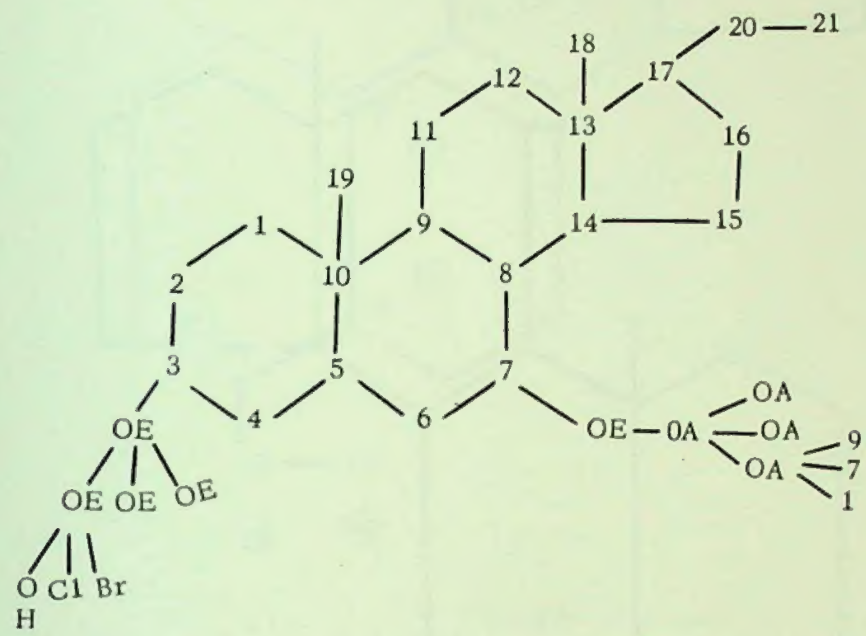


Figure 4.

Each carbon of the steroid nucleus is uniquely identified as, the No. 1 carbon, the No. 2 carbon and so on. For substituents on the nucleus, a special pattern is interposed between the substituent and the position of substitution. This pattern consists in the case of any substituent except an O-acyl, of an arrangement of pseudo elements designated as OE. In view of a limitation of the original program which restricts the maximum number of connections for any element to 4, the OEs were expanded as shown to accommodate a maximum of 9 substituents per position, which are, of course, possible only when they are alternatives for each other. Thus the three X variables are shown as alternative substituents in the "3" position. A search for an OH in the 3 position would ask for the chain 3 - OE - OE - OH. The computer would investigate all branches and determine whether a substituent could be found on this treelike pattern.

The same principle is used for the O-acyl but in view of the frequency of O-acyl substituents, a special OE - OA - OA - pattern is used to give a maximum of 9 - O-acyl substituents.

The compounds, both alternatives and equivalents, are set forth as composite formulas of this type. These composite formulas are assigned sequence numbers and coded in the same manner previously described. Both the preparation for coding and the coding were done by clerical, nonchemical personnel.

A file of 370 steroid patents was encoded this way but it must be emphasized that these 370 patents comprehend within their disclosures, by reason of the artificial genus practice, an estimated 3/4 million compound concepts each of which if requested either specifically or generically will be located within the file of patents.

Demonstration of Second Topological Search

This system is demonstrated despite the nonportability of the SEAC computer. By way of ordinary commercial communication circuits, a question is received here in Cleveland, where it is coded and the coding instructions are sent to the National Bureau of Standards in Washington, D. C. Mr. L. C. Ray of the National Bureau of Standards will take the instructions received as a punched paper teletype tape, feed them to SEAC and explain the searching operation.

The question was submitted by the school of Medicine of Western Reserve University in the form of a structural formula, illustrated in figure 5.

4 -0-ACYL, 11-KETO STEROID

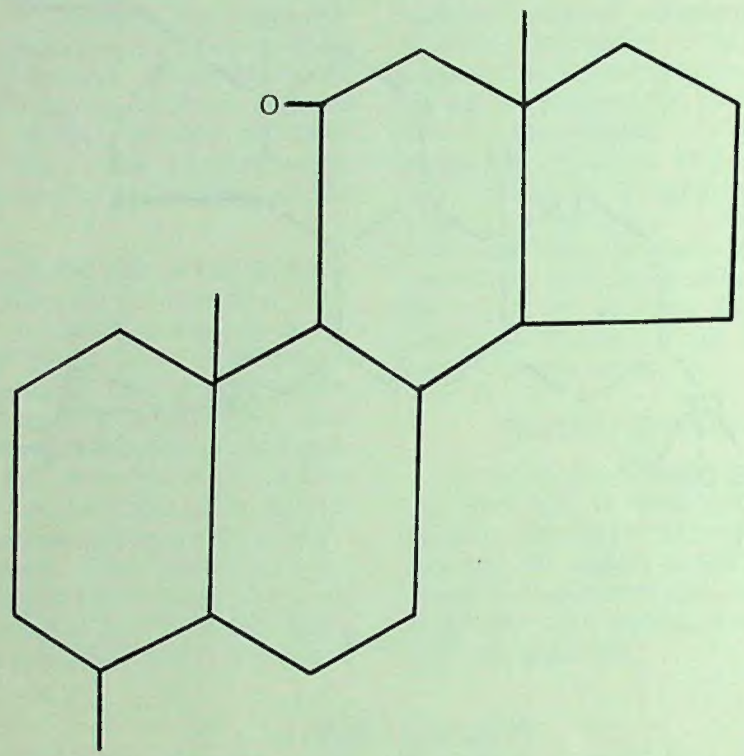


Figure 5.

The question specifies that a steroid compound be found where a keto (= 0) group is attached to position 11 of the nucleus and an O-acyl is attached at position 4. The question is silent as to the other substituents and double bonding patterns within the nucleus so it is assumed that the search will be satisfied by the finding of a steroid having at least the indicated structural requirements regardless of what else is present.

As will be noticed on the television screen, figure 6, a member of the Patent Office staff, Mr. Pfeffer, has sketched the steroid nucleus, identified the various positions and prepared it for coding as follows: He adds an OE-OA to the 4 position which indicates the requirement for any O-acyl group in the 4 position, and OE-OE to the 11 position followed by a double connection to oxygen, which is code 28. Figure 6

He now selects any path through the nucleus which shows the connectivity among the required groups. He then codes the question in the manner that has been described, as follows, figure 7.

This code has been punched on paper tape and is ready for transmittal to Washington, D. C. where the search will be performed.*

The answer is received in the form of the Patent Numbers 2,673,864 and 2,727,912.

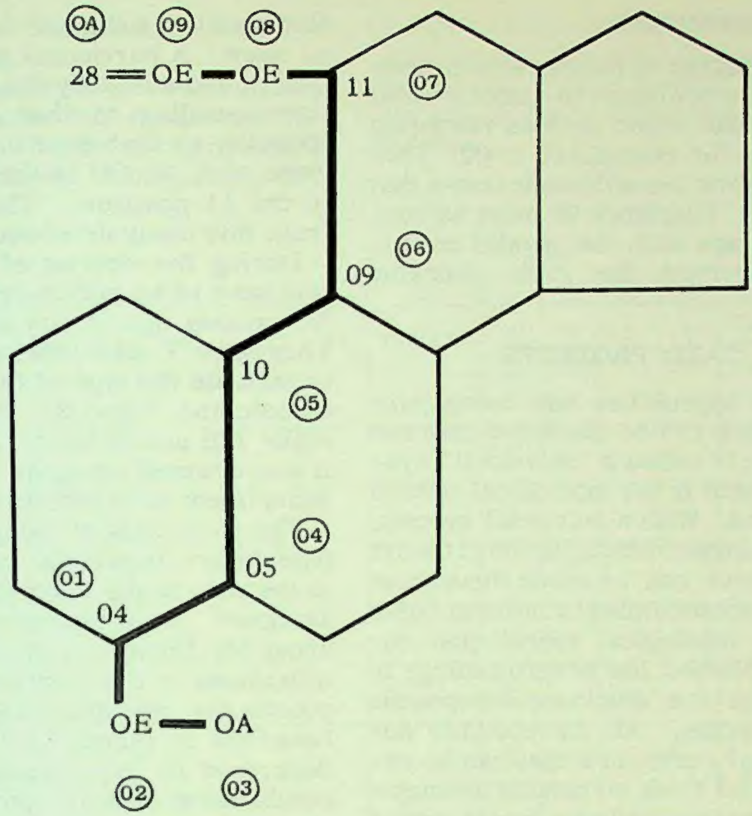


Figure 6.

No	Hex	IV	III	II	I	Element
1	01			02	04	04
2	02			03	01	0E
3	03				02	0A
4	04			05	01	05
5	05			06	04	10
6	06			07	05	09
7	07			08	06	11
8	08			09	07	0E
9	09		0A	0A	08	0E
10	0A			09	09	28
11	0B					

Figure 7.

Checking Routine Demonstration

SEAC has been instructed to locate certain types of coding errors. This operation is demonstrated by making an intentional error such as removing one of the connections, for example 03 to 02. This is recognized as an error since the code shows that 02 is connected to 03. Therefore 03 must be connected to 02. The tape with the invalid code is transmitted to demonstrate this error checking routine.*

PUNCHED CARD PROJECTS

The mechanization approaches now being pursued in the Patent Office can be classified into two categories. The first is called a "universal" system, an example of which is the topological search routine first described. With a universal system, no classification or compartmentalization of the art is necessary. A search can be made throughout the entire body of the art according to uniform rules and methods. In the topological search just described, no pre-established list of terminology is necessary to describe the disclosed compounds and provide search terms. All the possible descriptors for chemical compounds that can be expressed by and derived from structural arrangements of elements are available as search terms and all disclosures of compounds containing the required configuration within the molecule are retrievable for each search, regardless of the type of compounds involved.

This is in contrast with the "statistical" system, exemplified by the 1950 punched card experiment[1] and a current punched card project in the steroid art, which is described herein. In the statistical method a body of art is divided into practical and workable homogeneous portions. For each of these segments, a list of descriptors is established and each search is confined to the available descriptors and to that segment of art in which these descriptors are applicable and according to the rules established for it.

The illustration shows the analysis sheet for the steroids and the punched card portion transcribed from this sheet, as used in the steroid punched card project.

The descriptors consist of terms for chemical groups, such as double bond, OH and so on, for each of 21 positions of substitution pertaining to the steroid nucleus. The chemical descriptors are assigned to designated columns while the position numbers are assigned to particular rows. Since there are only 12 rows on the punch card, two columns are therefore alloted for each chemical descriptor. Thus a double bond in any position 1 thru 9 is indicated in column 1, rows 1 thru 9, respectively, and a double bond in any of positions 10 thru 21 is indicated in column 2, rows 0 thru 12, as seen. A particular group on a particular position is indicated by the intersection of the column corresponding to that group and the row corresponding to that position. Thus the sheet shows, inter alia, an OH in the 3 position and a keto (=0) in the 11 position. The card is directly punched from this analysis sheet, as indicated.

During the course of the analysis and coding it was seen to be necessary to add more terminology for greater specificity as to types of O-acyl groups. Therefore 9 additional such groups were provided to indicate the type of O-acyl indicated generically in columns 7 and 8. While this method does not make full use of the facilities of the punched card, it was deemed adequate for the initial stages of the experiment to determine practicability.

The thousands of compounds disclosed in the 370 patents are at present classified by one term, which is the title of the subclass to which the patents are assigned. By this punched card method 17 x 21 or about 350 terms are individually available as classifications of the compounds. In addition, the compounds are multiple coded, i.e., described by combinations of terms so that each compound can be described by any combination of these terms. A combination of descriptors constitutes, in effect, a synthesis of terminology.

The formulas encoded were what are called composite formulas. That is, each patent generally discloses a large number of compounds. There usually, is however, an equivalent or analogous relationship among them, the compounds having both a common configuration and a common utility. The composite formula, then, attempts to describe, in one formula, the concept of many equivalent formulas. Thus, it will be noted in the illustration, that the steroid has 5 substituents in the 5 position which is chemically impossible. Since such a search would not ordinarily be made no great problem is seen in this direction.

The composite formula device will select, among others, answers which are not entirely on all fours with the search request. For example, since a disclosure of both a 2 halo steroid and a 3 hydroxy steroid is coded in the same way by composite formula as a disclosure of a 2 halo, 3 hydroxy steroid, a search for one will retrieve the other. The actual desired compounds however, will be also retrieved and those beyond the scope of the question may be sufficiently analogous that the searcher may wish to see them too.

This system for the 370 steroid patents is in operation and to date about 42 applications for patent in the steroid art have been searched by punched card machine. Comparisons have been made with 30 of these applications searched manually by the conventional system and results are highly successful, no pertinent reference having been missed.

This initial procedure has been amplified to include more descriptors and give greater precision and more versatility as to generic search descriptors. Work meanwhile continues on coding the rest

---

[1]Mechanized Searching in the U. S. Patent Office, Bailey, M. F., Lanham, B. E., and Leibowitz, J. Published in Patent Office Society, 35, 566-587, Aug. 1953.

of the steroid art which, in total, contains about 2,000 patents.

It is important to emphasize, that the machine not only provides a rapid search but it also permits searching by many more terms than are provided in the conventional classification. The thousands of compounds disclosed in the coded patents are conventionally classified in one category, namely the subclass. By this method they are classifiable by any one or more terms selected from 168 individual terms.

2773888

		1	3	5	7	9	11	13	15	17	19	21	23	25
	Checked No.	=	OH	= O	Oacyl	Hal	α Allo	H	Oalkyl	V	W	X	Y	Z
0	22													
1	1	X								X				
2	2					X				X				
3	3	X	X	X					X		X			
4	4	X				X		X		X				
5	5	X					X			X				
6	6									X				
7	7													
8	8													
9	9													
10	10							X						
11	11			X										
12	12													
13	13													
14	14													
15	15													
16	16													
17	17		X				X	X						
18	18													
19	19													
20	20		X	X	X									
21	21		X		X			X						
		=	OH	= O	O Acyl	Hal	α Allo	H	O Alkyl	V	W	X	Y	Z
		2	4	6	8	10	12	14	16	18	20	22	24	26

Figure 8.

ILAS (THE INTERRELATED LOGIC ACCUMULATING SCANNER)

Preface

This is the second part of a set of two reports which embody the presentation made at the Western Reserve University Symposium for Systems on Information Retrieval held in Cleveland, Ohio on April 15-17, 1957

ILAS, (the Interrelated Logic Accumulating Scanner), was designed by personnel of the Office of Research and Development of the U. S. Patent Office and was constructed at the Bureau of the Census. We will describe a proposed system for use with ILAS and will then demonstrate a search making use of an existing scheme.

The contemplated system will use the topological coding principles previously described in the SEAC demonstration as much as is possible within the limits of a punched card operation. Chemical structures will be thought of as containing two types of building blocks, the ring configuration and the chain configuration. Each ring within the formula will be encoded in terms of the number of elements in the ring, the kind of elements, the number of each kind and the sequential arrangement of the elements. This will permit generic searching wherein, for example, a search for a nitrogen-containing heterocyclic ring compound will retrieve all nitrogen heterocyclic compounds, *or* a search for a compound having a nitrogen and 2 oxygens in 1, 3, 4 positional relationship will retrieve all compounds meeting these terms. For coding the chain configuration, several different approaches are under consideration, one of them being a so-called nodal method wherein each element is described in terms of its nearest neighbors. This does not give the precision of the topological method but offers promise of successful operation on the basis of statistical approximations.

There is an organizational format in a disclosure, such as the relationships among the various disclosed concepts that concern one compound, the relationships between the compounds in an admixture, the sequence of steps in a process and the various processes in the document. In addition, there is a variable number of each of these organizational units in each larger unit. This organization will be reflected in coding by the use of "grouping" signals which have the dual role of first, separating small units of disclosure, such as the several codes describing one compound from the several relating to another, and second, grouping together related units such as the several compounds in a mixture.

An important feature of the new system will be the extensive use of what is called the "interfix."[2] This is a grouping device for showing relationships among things which cut across any prearranged grouping organization. For example, in a paragraph of written material, a word or phrase in one sentence can be related to a word or phrase in another sentence to constitute a sentence not appearing as such in the context. This can be done by labeling the words so related with the same numerical value and adopting a rule that words having the same numbers are associated as being in the same sentence. This device can be used in many ways. For example, in the process $(A+B) \rightarrow (C+D)$ there are two separate compositions, each grouped by parentheses, namely (1) $(A+B)$ and (2) $(C+D)$. By writing $(A_1+B_1) \rightarrow (C_1+D)$ the same grouping arrangement is kept but A, B and C are further grouped by the interfix to indicate, in addition, that $A+B \rightarrow C$ and that D was later added.

A different meaning is given by writing $(A_1+B_1) \rightarrow (C_1+D_1)$ which signifies that the reaction products are $C+D$ rather than C alone. Note that both types of disclosure retain the same grouping arrangement.

In the structures, a major use of the interfix will be to link building blocks to each other as ring to ring and ring to chain.

The format of the contemplated punched card is shown in figure 9.

Standard 80 column IBM cards are used. Each code word is punched horizontally across the 80 positions of a single row of the card, as many as 12 codes being punched on a single card. As many cards as may be required for a document are used and scanning proceeds continuously from row to row and card to card.

Hexadecimal digits are used to represent the code. Four punching positions, therefore, are required for each code digit. Positions 1 to 4 in each row are used to designate grouping signals, positions 69 to 80 are for interfixes. When a ring is

---

[2] "Advances in Mechanization of Patent Searching,"--Lanham, B. E., Leibowitz, J., Koller, H. R.

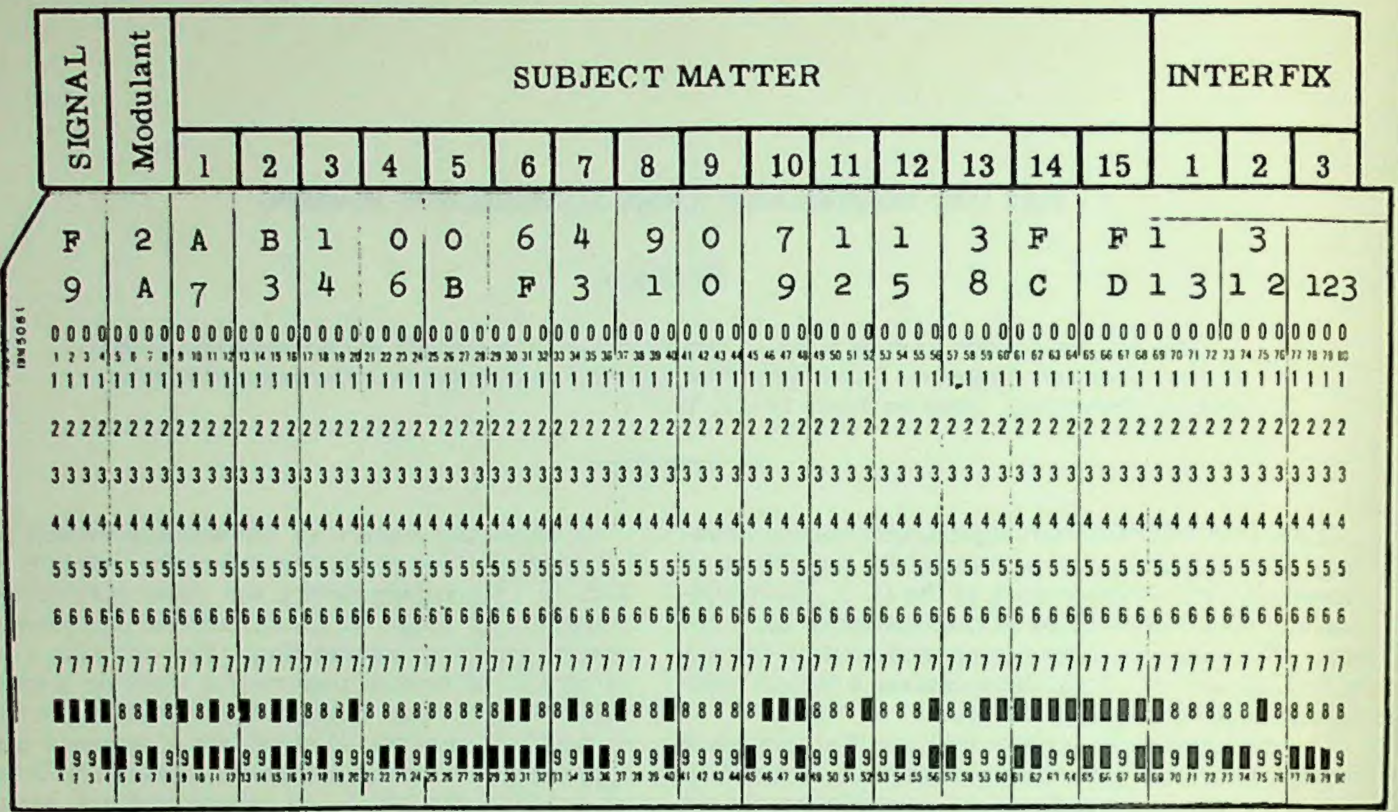


Figure 9.

attached to a chain, codes describing these two blocks will each have a punch in the same position in their interfix field. Positions 5 to 8 are used for what are called "modulants." These are used to modify or indicate the proper interpretation of the rest of the subject matter code. Thus if two subject matter codes each describe the same sequence of elements in a block, the modulant of one of them may indicate that the sequence described occurs in a ring and a different modulant in the other may indicate that it occurs in a chain.

ILAS is an 80 column sorting machine which is extremely flexible in concept. It is programmed in part by plug board wiring and in part by rotary switches, facilities for both of which appear on the console. The operator sets up the question to be searched by wiring the plug board for grouping signals and interfixes and sets a series of rotary switches for the subject matter code. As many as 12 subject matter codes can be included in a single question.

In the major timing cycle upon which the machine operates, corresponding to the interval for scanning one row, provision is made for detection and response to as many as 12 independent grouping signals. Upon finding any one of them, a test pulse is fed through certain "tentative circuits" which have been set up and if the appropriate information has been found, the test pulse can get through the complete circuit and activate a relay. This relay then becomes part of another circuit to be used by a test pulse triggered by recognition of another, higher order grouping signal. Upon occurrence of any test pulse, if the circuit is only partially complete, the relays in that circuit which had been activated are dropped out. For example, if 2 codes must be found in an "item" (the group of codes describing a single compound), when an "end of item" signal is found, if both of the desired codes have been found, as would be indicated by two "subject matter code" relays being in the activated state, the test pulse activates an "item relay" and the 2 subject matter code relays are dropped out. However, if only one of these codes has been found, the "subject matter code" relay representing that one will be dropped back to its normal position.

When the "end of document" signal is found, all lower order relays are dropped out and if an answer to the entire question has been found in that document, the last card relating to that document, as well as the subsequent nonselected cards, are sorted into the next pocket. When a further "hit" occurs, sorting begins in the next pocket, and so on. Document identification is then made by the operator inspecting the bottom card in each sort pocket, the identification being printed on the card.

The coding scheme described is still in the formative stage. However, ILAS can be demonstrated by using a deck of punched cards prepared according to a system developed and tested in the Patent Office in the spring of 1950.[3]

The subject matter of this early project was medicinal compositions, a composition being a physical admixture of two or more ingredients. The ingredients were chemical compounds and complex natural products such as various plant and animal extracts. In addition to the ingredients there were disclosures of uses, properties, physiologcal behavior and so on, such terms being called broadly "functions." Included as functions were such

---

[3]Mechanized Searching in the U. S. Patent Office, Bailey, M. F., Lanham, B. E., and Leibowitz, J. Published in Patent Office Society, 35, 566-587 Aug. 1953.

things as diseases treated, parts of the body affected and etiological factors of disease. The ingredients and functions were listed in a classification schedule showing generic and specific relationships. There were also provided a schedule of chemical compounds, a plant and animal schedule and a classification of diseases in terms of body sytems and causative agents.

In coding the compositions, each ingredient was assigned a number of codes indicating various characteristics of the ingredient such as structural groups, functions, source (if a natural product), and so on. The multiple codes for each ingredient were tied together as pertaining to the same ingredient by a grouping signal which indicated the end of the ingredient descriptors. The set of ingredients in a composition were tied together by another signal indicating the end of the composition.

In view of the indentation arrangement of the schedule terms, the code for each term contained within it the codes of all terms generic to it so that when a generic search was made, all the specific embodiments were automatically retrieved.

Each code was punched in a horizontal row of the card and was segmented into position fields to show the generic-specific indentation pattern. There was no fixed row assignment for any code and the scanning proceeded continuously from row to row and from card to card. There was, therefore, no limit to the number of codes per ingredient, the number of ingredients per composition or the number of compositions per patent.

The method gave great flexibility in availability of search terms. An ingredient could be asked for by any one or a combination of terms, each term being itself selective of all the specific embodiments embraced. Correlation could be made between structural formula groups and functions. In effect, the terms were synthesized as required for the search. Further versatility in building up terminology was achieved by use of negative directions such as "Find A + B minus C, or A + B minus anything else."

In the performance tests, the search cards for 441 patents containing 6,272 disclosure items characterized by a total of 18,650 descriptive terms were scanned in 4.5 minutes or at a rate of 95 patents per minute.

In the demonstration of ILAS, the question asks for a combination of

(1) Folic acid, the growth factor

(2) Liver Extract, and

(3) A Sulfonamide, this being asked for generically by fewer codes than are required to specify a particular member of this class of compounds. The entire mixture is to be disclosed for use in treating the teeth.

A patent is found which teaches the combination of (1) yeast, which contains folic acid, (2) liver extract, and (3) sulfanilamide, which is *a* sulfonamide. It is disclosed for use in the therapeutic treatment of bones and teeth.*

Many problems remain to be solved. We expect to continue the approaches we have described and to utilize all means at our disposal, to achieve our goal of an effective and rapid search system for the Patent Office.

---

*Presented before the Division of Chemical Literature, 129th meeting of the American Chemical Society, Dallas, Tex., April 11, 1956. Reprinted in U. S. Patent Office Society, 38, 820-838, Dec. 1956. Revised and printed as a Patent Office Research and Development Report.

37184--U.S.Dept.of Comm--DC--1957